
The Development of Second Language Writing Complexity in Groups and Individuals: A Longitudinal Learner Corpus Study

NINA VYATKINA

University of Kansas

Department of Germanic Languages and Literatures

1445 Jayhawk Boulevard

Room 2080

Lawrence, KS 66045

Email: vyatkina@ku.edu

This study explores the development of multiple dimensions of linguistic complexity in the writing of beginning learners of German both as a group and as individuals. The data come from an annotated, longitudinal learner corpus. The development of lexicogrammatical complexity is explored at 2 intersections: (a) between cross-sectional trendlines and the individual development paths of 2 focal learners and (b) between different complexity variables. The study contributes to the empirical body of linguistic complexity research by close tracking of beginning learners over 4 semesters of collegiate study of German as a second language (L2). For this purpose, data for multiple variables were collected at dense time intervals using multiple waves, and correlation analysis between various datasets was performed. The results confirm some general developmental trends established in previous research. However, the study also found significant variability between individual and cross-sectional data. Furthermore, differences found for more specific complexity measures between this study's results and previous research are explained in terms of differences in instructional approaches. In addition, the study contributes to the discussion of methods and metrics appropriate for tracking the development of complexity in foreign language writing. The study concludes with implications for L2 pedagogy and further research, including applications of computational methods.

THE NOTION OF TIME IS CENTRAL TO ALL disciplines concerned with human development, including second language (L2) studies. To make inferences about how learners develop over time, most L2 studies have applied either cross-sectional designs (where researchers compare data from different groups of learners at different proficiency levels) or "classical" longitudinal designs (with few waves of data collection from the same participants over a relatively long period). However, more recently, many prominent researchers have called

for employing longitudinal designs with dense developmental data collection (de Bot, Lowie, & Verspoor, 2011); describing the interaction between cross-sectional and longitudinal data (Byrnes, Maxim, & Norris, 2010; Larsen-Freeman, 2006); looking beyond stable developmental patterns and accounting for variation and degrees of this variation (Byrnes, 2009; Ortega & Byrnes, 2008; Pallotti, 2009); and capturing the interrelation of multiple developing interlanguage subsystems (Larsen-Freeman, 2009; Verspoor, de Bot, & Lowie, 2011; Verspoor, Lowie, & van Dijk, 2008).

This study responds to these calls by exploring the development of multiple dimensions of linguistic complexity in the writing of beginning learners of German both as a group and as

The Modern Language Journal, 96, 4, (2012)

DOI: 10.1111/j.1540-4781.2012.01401.x

0026-7902/12/576-598 \$1.50/0

© 2012 *The Modern Language Journal*

individuals. The data come from an annotated, longitudinal learner corpus, which comprises writing samples from American beginning college-level learners of German collected at regular short intervals over their first four semesters of study. The development of lexicogrammatical complexity is explored at two intersections: (a) between cross-sectional trendlines and the individual development paths of 2 focal learners and (b) between different complexity variables. Furthermore, the study contributes to the discussion of methods and metrics appropriate for tracking the development of complexity in foreign language writing.

The study is organized as follows. The next section presents the study background by first discussing the construct of linguistic complexity and then reviewing relevant research literature. The next sections report on the study. The description of the research purpose and questions is followed by a detailed methodology section. It begins with introducing the participants and then describes the instructional approach, design, measures, and computational procedures. The results section presents both cross-sectional and longitudinal data and reports correlations (a) of different complexity measures with time and (b) between measures. At the end of the results section, a qualitative collocation analysis is presented. The results section is followed by summary and discussion, including comparisons with previous research. The last section presents conclusions and implications for further research. Finally, the elicitation tasks used in the study are listed in the Appendix.

STUDY BACKGROUND

Linguistic Complexity as a Developmental Construct

Linguistic complexity is one of the components of the three-dimensional L2 proficiency model encompassing Complexity, Accuracy, and Fluency, or CAF (Skehan, 1989). According to contributors to a recent special issue of *Applied Linguistics* on CAF research, these measures have, since the late '70s, "figured as major research variables in applied linguistic research" (Housen & Kuiken, 2009, p. 461) and have been used for assessment of learner performance on specific written and oral tasks, as indicators of proficiency levels, and as milestones for learner progress. Although it is generally recognized that the three measures are closely interrelated, many studies focus on one specific dimension, notably complexity.

Complexity is defined as "[t]he extent to which the language produced in performing a task is elaborate and varied" (Ellis, 2003, p. 340) and may include both syntactic and lexical features (see Bulté & Housen, 2012 and Ortega, 2012 for an extended discussion). Research synthesis studies (Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998) have established that the average number of words per T-unit (main clause with all dependent subordinate clauses) has been the most frequently used syntactic complexity measure. In addition, researchers have used the indices of words per clause, clauses per T-unit, and dependent clauses per clause. In a recent comprehensive study that investigated a large set of complexity measures, Lu (2011) suggested that the clause and ratios of various features per clause are the best indicators of different L2 proficiency levels. Lu also advocates using more specific complexity measures in addition to general complexity measures for evaluating proficiency.

A number of complexity studies supplement length-based complexity measures with lexical density and diversity measures. Lexical density is typically operationalized as the ratio of lexical to functional or total words, and lexical diversity as the type-token ratio (TTR), or the ratio of different words to all words in a text (Polio, 2001, p. 99). Studies investigating different aspects of linguistic complexity are reviewed in the next sections.

Complexity Studies on Second Languages Other Than German

Most early studies found a general increase of average CAF values as learners progressed through instructional sequences (e.g., Arthur, 1979). However, all researchers also discovered considerable between-subject and within-subject variation. Larsen-Freeman (1983) explored both written and oral English as a second language (ESL) learner productions in a cross-sectional and a longitudinal study and found that the frequency of T-units and error-free T-units could be used as indicators of proficiency levels but varied depending on the language mode and task. There was also considerable variation between individuals. Arthur also discovered significant intra-individual variation in the development of ESL writers over 8 weeks on all CAF measures (accuracy, spelling accuracy, and length), even in cases with a general monotonic group-level developmental trend. Casanave (1994) came to a similar conclusion exploring English as a foreign language (EFL) writing of Japanese students and strongly suggested that development be studied for individual

learners rather than based on group averages. Furthermore, Arthur and Kern and Schultz (1992) showed that the T-unit length increases for beginning learners but, at a certain point in the instructional sequence, it starts to decrease. The authors attributed this finding to the fact that more proficient learners use more embedding rather than longer syntactic constituents. These results from early research have been fine-tuned by more recent longitudinal studies, which are reviewed below.

A series of recent studies written from a Dynamic Systems Theory (DST) perspective sparked renewed interest in CAF in general and complexity in particular (e.g., Ellis & Larsen-Freeman, 2006; Verspoor et al., 2011). Researchers who work in this paradigm are primarily interested in “intra-individual and inter-individual variation over time” (de Bot et al., 2011, p. 2). For example, Larsen-Freeman (2006), using a time-series design (four observations over 10 months), showed that while all CAF group averages of her ESL learners consistently increased over time, the participants exhibited notable developmental variability as individuals. As far as complexity is concerned, one participant showed considerable development in lexical complexity while lagging in grammatical complexity, whereas another participant followed an opposite developmental pattern. Finally, Larsen-Freeman showed how a qualitative analysis of learner writing sheds additional light on the participants’ developmental profiles and recommended adding a qualitative component to future developmental studies.

Verspoor et al. (2008) and Spoelman and Verspoor (2010) explored the dynamic interaction of different writing complexity variables in longitudinal case studies. The former study described the development of an advanced EFL learner, and the latter study focused on beginning stages of L2 proficiency in Finnish. Writing samples were collected in both studies at dense time intervals over a 3-year period, in which both lexical and syntactic complexity features were explored. The analysis showed a complex relation between these variables that changed dynamically over time. For example, in the first study, during the first observation points, varied word use (TTR) and length of sentence (SL) showed a positive correlation, thus acting as “connected growers” (van Geert, 1994). However, from observation 4 to 15, these measures showed a negative correlation and, therefore, a competitive relationship to each other. The authors of both studies concluded that some complexity features develop hand in hand by using the same attentional resources, whereas

others may require the full attention of the learner during his or her active development.

Complexity Studies on L2 German

An early investigation of complexity in L2 German is Cooper (1976). Cooper compared a number of syntactic complexity measures across five proficiency levels, including four levels of college-level L2 writers and also professional native-speaker newspaper writers. He found that clause length, subordination ratio, T-unit length, sentence length, as well as the number of certain sentence-embedding constructions steadily increased with each adjacent level (roughly equivalent to 1 year of study), but the increase was significant only with every second level. It should be noted that Cooper only explored cross-sectional data from 10 participants at each level and did not consider any longitudinal data. The following sections review more recent L2 German complexity studies that explicitly focused on longitudinal research methods and data.

The authors in many studies reviewed in the previous section advanced the point that at least some syntactic complexity measures not only cannot be strictly separated from, but are inextricably intertwined with, lexical complexity measures. As Ravid (2005) has argued, “[c]ause length derives from number and length of intraclausal phrases, which in turn reflect lexical density and diversity, combined with syntactic depth and diversity” (p. 351). This approach reflects the premise of Systemic-Functional Linguistics, which postulates *lexicogrammar* as a sole complex object of linguistic inquiry instead of the two separate layers of grammar and lexicon (e.g., Halliday & Matthiessen, 1999).

Following the SFL approach, the research team working on the curriculum project Developing Multiple Literacies at the Georgetown University German Department (GUGD) has published a series of studies exploring essays of developing L2 German student writers with the focus on various lexicogrammatical patterns as “forms of textual meaning-making” (Byrnes et al., 2010, p. 38). In contrast to most studies on L2 grammatical development, which rarely provide details of the respective pedagogical settings, the overarching goal of these studies has been to explore linguistic features of student writing as specific learning outcomes achieved in response to a carefully designed research-based 4-year-long curriculum implemented at GUGD.

In a longitudinal study, Byrnes (2009) explored the development of subclausal complexification in

terms of nominalization and grammatical metaphor (GM) in student writing. She found that the frequency of these features increased dramatically from level (year) III to level IV, which directly reflected the shift of the instructional focus from more verbal (narrative) to more nominal (expository) text genres and writing styles, and thus demonstrated a desired learning outcome. Furthermore, the study showed that, whereas lexical density (measured in content words per clause) increased significantly over three instructional levels, grammatical intricacy (measured in clauses per sentence) slightly (yet insignificantly) decreased. Another important finding, facilitated by the longitudinal design, is that students who seem average based on more general complexity measures such as clause length may demonstrate speedier development of more specific measures such as the amount of nominalization (see also Byrnes & Sinicropo, 2008, for the case of relativization). Ryshina-Pankova (2010) expanded this line of research by exploring the use of the grammatical metaphor as a means of thematic progression in L2 German written texts. She identified different GM types as characteristic of various (advanced) acquisition levels and demonstrated how GM use contributed to greater or lesser communicative success of learner texts.

Byrnes et al. (2010) is a recent comprehensive study from this series, which investigated, among other targets, the development of several syntactic complexity measures. The results showed that general complexity (words per T-unit) increased incrementally over four curricular levels, whereas more specific measures exhibited more complex developmental patterns. The authors provided a curricular explanation for the latter finding. The number of clauses per T-unit increased significantly in level II due to the instructional focus on narration, which is characterized by various forms of subordination. This trend continued in level III but was also paired with a significant increase in subclausal complexification (words per clause) due to the added instructional focus on public discourses which are characterized by extensive nominal structures and, therefore, longer clauses. Finally, clauses became again significantly longer in level IV but the amount of subordination decreased. The latter result confirmed Byrnes's (2009) finding that showed increased clause length due to extensive use of nominalization by the same learner population in response to the exclusive instructional focus on secondary, public discourses in level IV. Furthermore, this finding corroborated results from earlier complexity studies that found leveling of subordination but increases in clause length at more

advanced proficiency levels (see Ortega, 2003, for a review).

Importantly, the developmental patterns found by Byrnes et al. (2010) for "course-embedded [...] Prototypical Performance Tasks" or PPTs (p. 163), were confirmed when tested cross-sectionally using the same "Baseline Writing Task," or BWT, at all curricular levels. However, the differences between levels were more distinct when measured in response to PPTs, which led the authors to conclude that curricular-embedded tasks are more conducive to investigation of syntactic development. Finally, Byrnes et al. showed that their L2 German undergraduate students achieved similar levels of grammatical complexity as graduate students in Cooper's (1976) study or even surpassed them, which the authors again attributed to positive learning outcomes of the curricular innovation consistently implemented in the department's pedagogical practices.

RESEARCH PURPOSE AND QUESTIONS

The present study will contribute to the empirical body of linguistic complexity research by closely tracking learners from the beginner level and over four semesters of collegiate L2 German study and by comparing cross-sectional cohort data and longitudinal data for 2 individuals. For achieving this purpose, data for multiple variables were collected at dense time intervals using multiple waves, and correlation analyses between various datasets were performed.

The study aims to answer the following research questions:

1. How does L2 German writing complexity develop cross-sectionally for a cohort of learners and in 2 individual learners?
2. What is the relationship between the developmental paths of the cohort and the individuals?
3. How do developmental paths measured by different complexity metrics correlate with each other?

METHOD

Participants

The data for this study were collected from students enrolled in beginning and intermediate German courses at The University of Kansas during four sequential 16-week-long semesters, specifically in the first semester course in the spring of 2008 (80 hours of instruction); second semester, fall 2008 (80 hours of instruction); third

semester, spring 2009 (48 hours of instruction); and fourth semester, fall 2009 (48 hours of instruction). Most students were completing their four-semester-long language requirement and had little or no knowledge of, or exposure to, German prior to their enrollment. Also, an overwhelming majority of the students grew up in the Midwestern region of the United States and had American English as their first language (L1). It remained a relatively homogenous group of learners in terms of type and amount of exposure to the target language; that is, it was mostly limited to classroom interactions and instructional materials.

Cohort. Although the data were collected during consecutive semesters, the actual participant constituency changed from semester to semester as students withdrew from the program or joined it at some later point via placement test. Furthermore, data were collected only from those who had signed consent forms at the beginning of each data collection semester, and not all students chose to participate each semester. Therefore, writing samples from the described student population were used as cross-sectional cohort data for calculating group averages rather than longitudinal data in this study.

Individuals. Furthermore, the study focuses on 2 individual learners from the same population for longitudinal data collection, who were assigned the pseudonyms "Braden" and "Cassie." These 2 participants (from a total of 7 who completed the full four-semester-long sequence) were selected because they had a number of similarities in their language learning history. Both of them were "true beginners," that is, they had not studied German before and had never visited German-speaking countries. Both participants were not majoring in German, lived all their lives in the U.S. Midwest, and had American English as their L1. Both of them had some knowledge of Spanish, and Cassie also had some knowledge of sign language. However, they were different in terms of gender and age: Braden was a 19-year-old male and Cassie was a female in her early 30s. Despite these differences, according to their instructors, both participants had some similarities in learning style: They were generally characterized as diligent learners who tried hard and were better at writing than speaking.

Instructional Approach

The students in this study were enrolled in a multisection beginning and intermediate German language program, in which all courses were taught by graduate student instructors under the supervi-

sion of the researcher. All instructors followed syllabi and textbooks that were uniform for each instructional level although they had freedom to design specific lesson plans following the coordinator's guidelines. Each course included a combination of spoken interaction, grammar explanation and practice, writing assignments, searching German Web sites, and creative culture projects. The writing component was allocated 20% of the curriculum and, respectively, of the total course grade at each level. This aspect of the curriculum design reflected the coordinator's conviction that "writing deserves systematic and continued attention in the foreign language classroom in its own right, not merely as a support skill for listening, reading or speaking" (Abrams, 2010, Conclusion section, para. 1; see also Byrnes et al., 2010). While a multidraft process writing approach was used at all levels, only rough drafts of written assignment are used as data in this study.

Tasks and Timeline

Students wrote their essays in response to level-appropriate writing tasks with topics and prompts reflecting relevant instructional content. In this regard, tasks in this study can be considered *Prototypical Performance Tasks*, or "curricular-level-specific PPTs" as defined by Byrnes et al. (2010, p. 179). However, the design of this study differs from the cited study in observation density and includes 18 PPTs collected approximately every 3 weeks as opposed to the four PPTs collected at the end of an instructional level each by Byrnes et al. The final task was a BWT (described in the next section).

The curricular progression of the genres of the writing tasks (see also Table 1 and Appendix) was as follows. The tasks at time points T1–T5 (first semester) and T6–T10 (second semester) required learners to write personal narratives. According to Maxim (2011), this genre involves "exploring self-identity in the German-speaking world through different roles that young adults assume in society (e.g., student, hobbyist, consumer, traveler, family member, citizen)" (p. 12) and is level-appropriate for first-year college-level language learners. T11–T14 tasks (3rd semester) also required students to write personal narratives and personal accounts but with added reasoning elements, which is appropriate for second-year language learners (Maxim, 2011, see also Byrnes et al., 2010). T15–T18 tasks (fourth semester) invited students to reflect on the meaning of the stories from the course book and to provide explanations of selected aspects and arguments supporting their opinions, which is appropriate for more advanced students. The final (T19) task asked

TABLE 1
Timeline, Tasks, and Data Pool

Semester	Time Point (T)	Samples	Words (Total)	Words (Mean)	Timed/ Untimed	Tasks From:	Chapter
First	1	28	1914	68	t	Briggs et al. (2008)	1
	2	25	2176	87	t		2
	3	26	3404	131	t		3
	4	27	2864	106	t		4
	5	25	2445	98	t		5
Second	6	40	3854	96	t	Teichert and Teichert (2005) BWT	6
	7	40	4164	104	t		7
	8	29	3056	105	t		8
	9	38	4083	107	t		9
	10	35	4072	116	t		10
Third	11	30	4148	138	t		11
	12	24	3627	151	t		12
	13	24	3517	147	t		13
	14	20	2758	138	t		14
Fourth	15	21	4550	217	u		1
	16	16	3907	244	u		3
	17	16	3780	236	u		4
	18	18	4276	238	u		10
	19	22	3547	161	t		NA

students to summarize the contents of a book of their choosing and to explain why they wished to recommend it to their German peers. This task thus combined elements of narration and argumentation.

Topic and Prompt Sources

In the first three semesters, the prompts were taken from journal writing assignments at the end of each of the 14 chapters of the course workbook (Briggs, Di Donato, Clyde, & Vansant, 2008), which were based on the material covered in the corresponding course textbook (Di Donato, Clyde, & Vansant, 2008). In the fourth semester, selected chapters from Teichert and Teichert (2005), a textbook based on short stories written by German-speaking authors, were used as primary course texts, and writing tasks were taken from those chapters. Finally, the T19 task was a slightly modified BWT, adopted from the GUGD Multiple Literacies project (see Byrnes et al., 2010), which is a generic writing task not directly related to instructional materials used in this study’s program.¹ The full list of writing topics is provided in the Appendix, and specific prompts can be found in the sources cited in this paragraph.

Writing Conditions

Collected writing samples were written under the following conditions. During the first three

semesters, students typed each of their essays in a computer lab during a 50-minute-long class period and submitted them through the online courseware (Blackboard). Make-up essays of students who were absent during designated writing sessions were not included in the analysis. Learners were required to write during the whole class period. In this way, they were writing under controlled, timed conditions, with the instructor’s supervision, and were allowed to consult the instructor and online dictionaries but neither online translators nor their textbooks or notebooks. In the fourth semester, the essay-writing procedure changed. Students wrote four out-of-class essays in response to textbook prompts (to allow more time for reflection requested in the tasks) and submitted them electronically via Blackboard on designated due dates. They were required to write at least one double-spaced page. At the end of the fourth semester, students wrote their final essay under the same conditions as timed essays in preceding semesters (during a 50-minute-long class period). In sum, essays were collected at 19 data collection points (T1–T19) during the four focal semesters. T1–T14 and T19 data thus represent timed in-class writing, whereas T15–T18 data represent untimed out-of-class writing (Table 1).

Sampling

With regard to the sampling procedure, the study followed the “instruction-embedded total-

sampling approach" (Byrnes et al., 2010, p. 165); that is, all essays written by the students of the focal cohort in response to instructional tasks were collected. It must be noted that not all participants turned in essays at all data collection points, including the 2 longitudinal participants: Braden did not submit essays at T11, T14, and T16, and Cassie at T8. Due to this fact and to participant attrition (see the section above), the number of writing samples that served as cross-sectional data in this study varied from time point to time point and ranged from 16 to 40 (Table 1). Table 1 also lists the total and average word count produced by the cohort at each time point.²

Measures

For choosing measures of linguistic complexity, this study relied on recent recommendations that Norris and Ortega (2009) derived from a comprehensive research review and synthesis. The authors strongly advise researchers to measure complexity multidimensionally and to supplement general measures such as sentence length with distinct, complementary, and specific complexity measures (see also Lu, 2011). Norris and Ortega identify the following main dimensions of syntactic complexity, the metrics for which should be chosen depending on specific conditions of each study: (a) overall or general complexity; (b) subclausal complexity; (c) complexity via subordination and coordination; and (d) variety, sophistication, and acquisition timing of forms produced (pp. 561–562).

In this study, a number of measures were used to tap into these multiple dimensions of complexity. In choosing from the plethora of available measures, it was decided to only use measures that allowed automatic annotation and/or searches for target features so as to eliminate as much subjectivity as possible in data tagging and calculations. As Granger, Kraif, Ponton, Antoniadis, and Zampa (2007) note, the value of a learner corpus increases exponentially for both theoretical and applied second language acquisition (SLA) research purposes when it is annotated for parts of speech (POS) and other linguistic properties and analyzed with natural language processing (NLP) tools. Still, despite the wide availability of computational resources, these applications remain largely underexploited in complexity research (see, however, Hawkins & Buttery, 2010; Saville, 2010), especially for languages other than English. This study aimed to benefit from available NLP resources and used automatic corpus tools for computing length-based measures as well as automatically assigned

POS tags as proxy measures for surface syntactic structures (Aarts & Granger, 1998). All metrics used in this study are ratios.

General Complexity. First, *sentence length* (SL, see Verspoor et al., 2008) was measured in the number of words per sentence (W/S). SL was chosen as a generic metric "with a potentially multiple-clausal unit of production in the denominator" measuring overall syntactic complexity (Norris & Ortega, 2009, p. 561). The sentence was chosen as the main unit of analysis instead of the widely used T-unit (Hunt, 1965) because SL can be calculated automatically, whereas T-units need to be coded manually. Furthermore, as Bardovi-Harlig (1992) argues, a T-unit analysis "artificially divides sentences that were intended to be units by the language learner, imposing uniformity of length and complexity on output that is not present in the original language sample" (p. 391). In contrast, by using the sentence, "the unit directly produced by the learner" (Bardovi-Harlig, 1992, p. 391), the researcher takes into account the learner's (conscious or unconscious) choice. Finally, since a T-unit analysis treats conjoined clauses as independent clauses, it "discounts the learner's knowledge of coordination" (Bardovi-Harlig, 1992, p. 391) which is an important indicator of complexity at beginning levels of language proficiency (see also Casanave, 1994; Ishikawa, 1995).

Clausal Complexity. The SL measure was supplemented by two specific complexity indicators. Sentence length can be increased by two different types of complexification: adding more coordinate or subordinate clauses to a matrix clause or making clauses longer (subclausal complexification). Thus, the first metric can be expressed in mean sentence length in clauses (clauses/sentences), and the second metric in mean clause length in words (words/clauses). It must be noted that, on the one hand, although clauses have been used as a unit of analysis in an overwhelming majority of complexity studies, they have not been defined consistently,³ which may engender differences in annotation and subsequent results (Bulté & Housen, 2012; Ishikawa, 1995; Lu, 2011; Polio, 2001). On the other hand, most CAF studies follow Hunt (1965), who defined a clause as "a visible subject and a finite verb" (p. 29). If clauses are restricted to finite clause units, the number of finite verbs can be used as a proxy for counting the number of clauses. The only difference between these two units of analysis would be in counting units constituted by coordinated finite verbs as distinct finite verb units (called FV-units by Verspoor et al., 2008), when they actually belong

to one and the same clause unit and have one and the same subject. However, the benefit of using FV-units instead of clauses is that the former are less ambiguous and can be found and computed by automatic POS tagging tools. Since only automatically computed measures were used in this study, subclausal complexity was measured by the ratio of words per FV-unit (W/FV) following Verspoor et al. To characterize complexification by coordination and subordination combined, the ratio FV-units per sentence (FV/S) was used.

Coordination and Subordination. To distinguish between complexification by coordination and subordination, two additional proxy measures were used, namely *coordinating conjunctions* (CC) and *subordinating conjunctions* (SC). Normalized CC and SC frequencies per 100 words were used to compare essays of different length. Coordinating conjunctions (e.g., *und* ‘and,’ *aber* ‘but,’ *oder* ‘or’) are connectors that combine homogenous syntactic constituents, or “parts that have ‘equal’ syntactic value” (Verspoor & Sauter, 2000, p. 101). They can connect coordinated subjects, predicates, or other intraclausal constituents as well as clauses. Coordination was thus explored as a broader syntactic construct not only limited to clauses. Subordinating conjunctions (e.g., *dass* ‘that,’ *weil* ‘because,’ *wenn* ‘when’) are syntactic connectors that function at the interclausal level, combining a main clause and its dependent clauses. Subordination was explored only in terms of adverbial and nominal clauses but not modifier clauses (expressed by relative pronouns and relative clauses).⁴

Lexicogrammatical Variety. Finally, lexicogrammatical variety was measured by a type–token ratio following some other studies that compared syntactic and lexical complexity (Larsen-Freeman, 2006; Verspoor et al., 2008). Namely, *corrected type–token ratio* (CTTR), also known as adjusted or sophisticated TTR, was used. CTTR is calculated in word types divided by the square root of 2 times the total number of words and thus “takes the length of the sample into account to avoid the problem that regular type–token ratios are affected by length” (Larsen-Freeman, 2006, p. 597, see also Lu, 2012). In addition, a qualitative type–token analysis of constructions containing coordinating and subordinating conjunctions for the 2 focal learners was undertaken to arrive at a more specific level of description of lexicogrammatical variety.

Computational Procedures

The units *sentence* and *word* were defined here based solely on typographic features. A sentence

was defined as “a string of words with a capital letter at the beginning of the first word and a period or another terminal punctuation mark after the last word” (Homburg, 1984, pp. 91–92). A word was defined as a string of letters separated by spaces from adjacent strings of letters. Sentence, word type, and word token frequencies were computed automatically using WordSmith Tools (Scott, 2008). Finite verbs as well as coordinating and subordinating conjunctions were automatically annotated in the learner corpus using the POS Tree Tagger for German (Schmid, 1994).⁵ Tag frequencies were then automatically computed using the WordSmith Tools *WordList* function (Scott, 2008). Computed frequencies were entered into a Microsoft Excel spreadsheet that was used for calculating finite verb frequency (a sum of finite full, auxiliary, and modal verbs) as well as ratios of words per sentence, finite verbs per sentence, words per finite verb unit, coordinating and subordinating conjunctions per 100 words, and CTTR. Next, graphs were created to illustrate the developmental course for each focal feature. Finally, the WordSmith Tools *Concord* function (Scott, 2008) was used for a qualitative analysis of CC and SC collocations (see Results II).

Cross-sectional and longitudinal data for all focal variables were plotted in Microsoft Excel, and a correlation analysis was performed. First, cross-sectional averages for the cohort were computed for each measure at each time point, and developmental trendlines for the cohort were analyzed. Next, it was established how the 2 focal longitudinal learners fared vis-à-vis this class average at each time point, and their developmental pathways were explored.

RESULTS I: CORRELATIONS OF COMPLEXITY MEASURES WITH TIME

Table 2 presents the correlation values between time and the six complexity measures for the cohort averages as well as for the 2 focal learners.

The following sections report on both cross-sectional and longitudinal results. Each section contains three graphs to illustrate how the cohort and the 2 focal individuals developed over 19 time points vis-à-vis a complexity measure. Chart (a) represents the quantified development of the cohort, with the average of each measure for each assignment and a regression trendline for the average calculated by least squares. To illustrate graphically the relative certainty of each of these trendline estimates, curved lines representing a 95% confidence interval of the predicted average are plotted, as well (Draper & Smith, 2001). Charts

TABLE 2
Correlations Between Time and Complexity Measures

	SL (W/S)	FV/S	W/FV	CC	SC	CTTR
Cohort	0.94***	0.82***	0.44	-0.63**	0.84***	0.89***
Braden	0.76***	0.86***	-0.29	0.07	0.66**	0.59*
Cassie	0.74***	0.54*	0.33	-0.18	0.62**	0.60**

Note. CC = coordinating conjunctions; CTTR = corrected type-token ratio; SC = subordinating conjunctions; SL = sentence length; W/FV = words per finite verb unit; W/S = words per sentence.
* $p < 0.05$, ** $p < .01$, *** $p < 0.001$.

(b) and (c) then compare individual development to the cohort average trendline. Braden and Cassie’s values for each measure and time point are represented in scatterplots, along with least squares regression trendlines and 95% confidence intervals for the trendline. These are superimposed on the cohort trendline for comparison purposes.

Sentence Length in Words (SL)

General syntactic complexity was measured by mean sentence length (number of words per sentence). Figure 1a shows that the class average sentence length increases with the instructional progression with very slight oscillations of the ascending trend, namely from 6 words per sentence at T1 to 10 words per sentence at T19. The latter result is similar to Cooper’s (1976), who found that on average, American students of German write 10.3 words per sentence in their second year of study.⁶ Figure 1a shows a fairly smooth diagonal increase with slight upward and downward oscillations. An extremely strong positive correlation between SL and time was found for the cohort ($r = 0.94$).

Figures 1b and 1c show that Braden and Cassie’s sentence lengths develop in a very similar way: The cohort trendline runs right through the center of Braden and Cassie’s trendlines. In other words, on this measure they can be taken as very good representatives of the average for this cohort. Respectively, sentence lengths of both focal learners strongly and positively correlate with time ($r = 0.76$ for Braden, $r = 0.74$ for Cassie). In terms of individual variation, Braden is more frequently above the class average in the first semester, below the average in the second semester, and around the average in the fourth semester. Cassie is either below or slightly above the average, with the largest fluctuations at T16 and T18. In the last semester, her sentence length is mostly above the class mean. This is reflected in

her trendline, which has a slightly higher slope than the cohort average (Figure 1c).

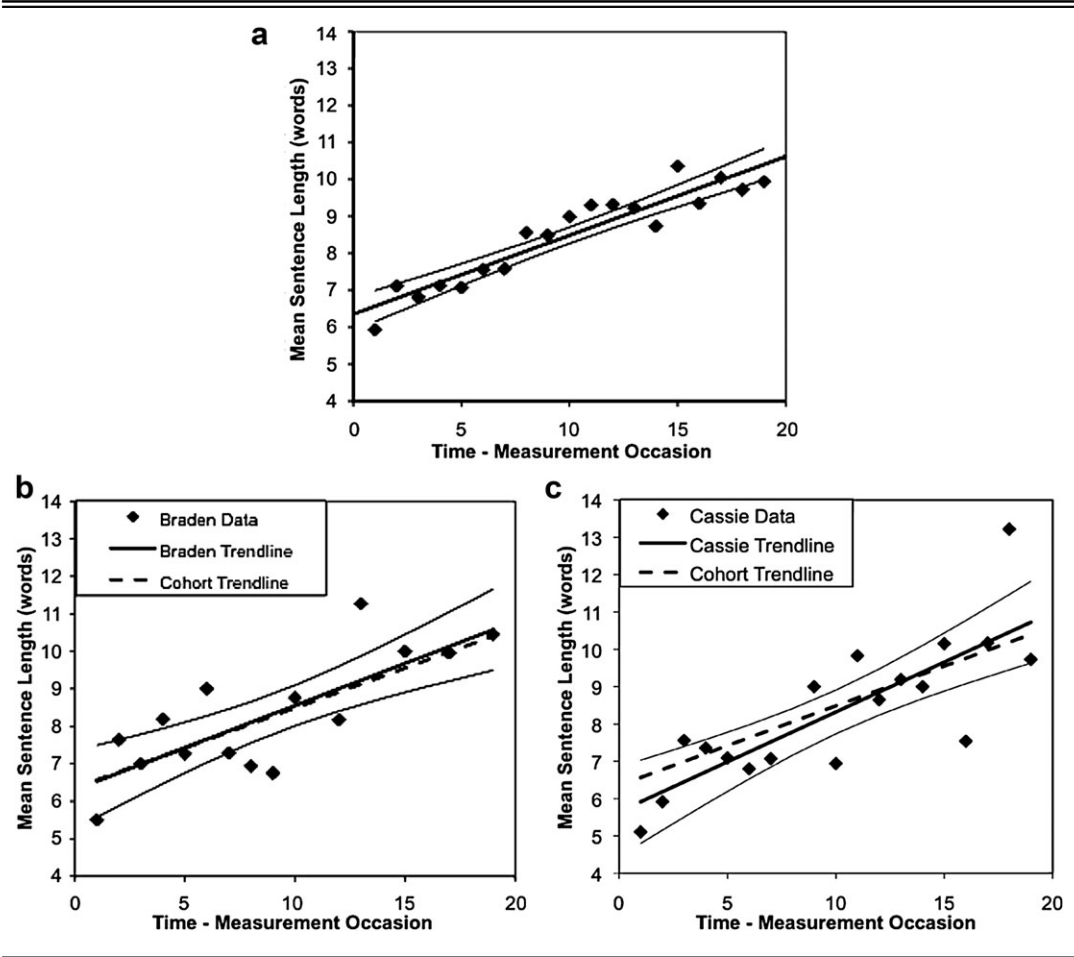
FV-Units per Sentence (FV/S)

The cross-sectional cohort data analysis (Figure 2a) shows a general increase of the average sentence length in FV-units. This means that learners use more finite verbs per sentence over time: The trendline goes up from 1.1 to 1.4 FV/S. This overall increase may not seem large; however, Ortega (2003) has shown that an increase for a similar measure (clauses/T-unit) is usually significant at a 0.2 level. The correlation between FV-units per sentence and time is positive and very strong in this data ($r = 0.82$).

The longitudinal data for both Braden and Cassie also show a general linear increase of FV-units per sentence and a positive correlation of this measure with time. For Braden, this correlation is very strong ($r = 0.86$) and for Cassie, it is more moderate but still significant ($r = 0.54$; $p < 0.05$). Figure 2b illustrates that Braden starts around the class mean, but after that, his FV/S ratio is consistently at or above the average (except T7 and T9) and surpasses it by a large margin of 0.43 at T19. Figure 2b shows a steeper slope for Braden’s trendline in comparison with the cohort trendline, which runs below Braden’s confidence intervals approximately after T12.

Cassie’s path follows Braden’s almost identically from T1 through T6 (her values staying just below Braden’s), but her later values are consistently lower than Braden’s and the class mean, with a few exceptions. Also, Cassie’s values are scattered more around the trendline, which shows a larger variation. Notably, Cassie’s FV/S value drops sharply down to the starting point (around 1 FV/S) at the final time point (T19). Figure 2c shows that Cassie’s trendline is below the cohort trendline and does not intersect it, with the slopes of both lines increasing almost in parallel over time.

FIGURE 1
Sentence Length Data and Trendlines: (a) Cohort Averages; (b) Braden With Cohort Trendline; (c) Cassie With Cohort Trendline



In sum, the cohort as well as the 2 focal learners use more FV-units per sentence over time. However, Braden's increase surpasses the class average, whereas Cassie's values, staying mostly below the class average, increase roughly at the same rate as the latter. Also, Cassie's data show more upward and downward oscillations, which moderates the correlation of the focal measure with time.

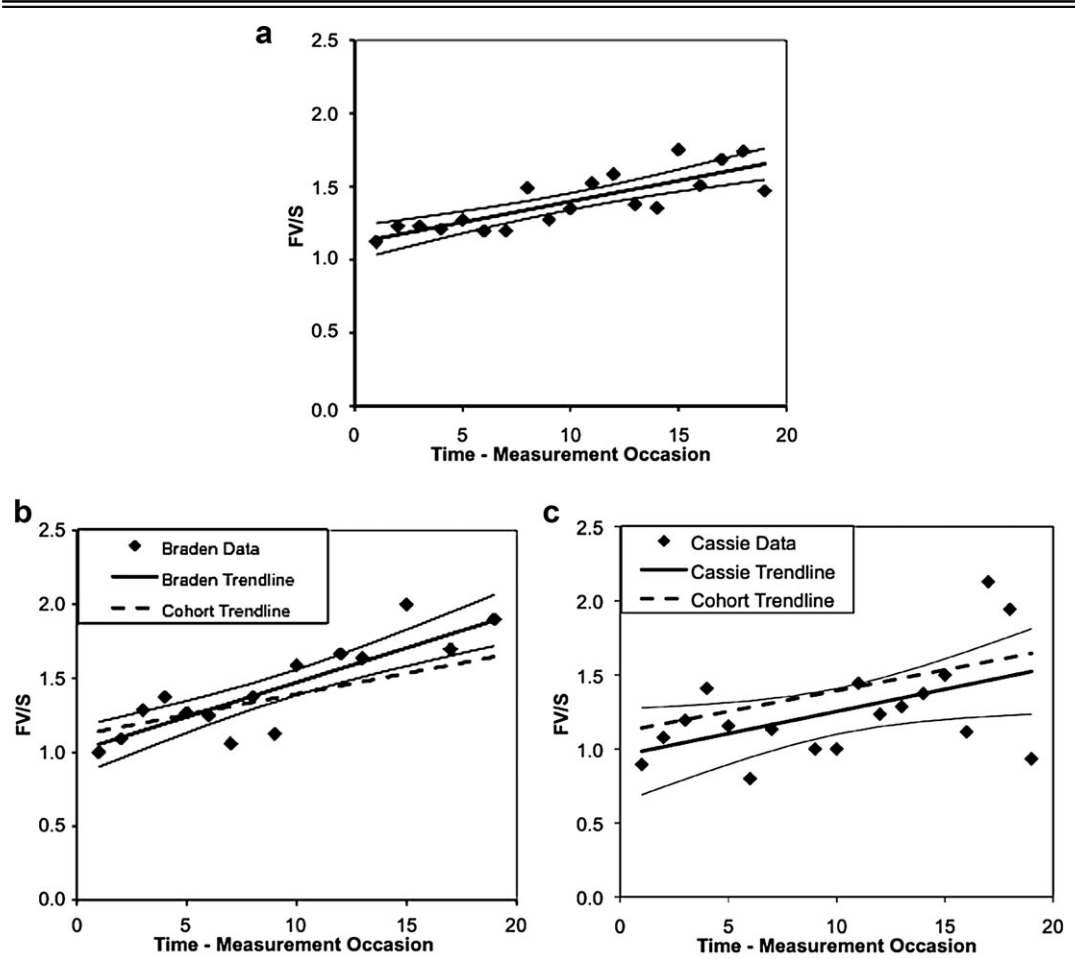
Words per FV-Unit (W/FV)

As opposed to the measures discussed above, the average length of FV-units in words did not show an obvious monotonic increase for the cohort data (Figure 3a): The values oscillate between 5.75 and 7 with no obvious pattern. The correlation analysis, however, showed a moderate positive correlation of W/FV measure with time ($r = 0.44$) that approached significance ($p = 0.06$).

The scattergram of Braden's data (Figure 3b) represents no discernable pattern between the length of FV-units and time, with W/FV values ranging randomly from 4.9 to 7.3. There is an indication of a slight negative correlation with time ($r = -0.29$), which is illustrated by an almost horizontal yet slightly declining trendline in Figure 3b. In other words, Braden does not begin to write longer FV-units over time; his FV-units show a trend of becoming shorter in contrast to the slightly increasing cohort average. However, this correlation is not significant ($p > 0.05$). Additionally, the differences between Braden's trendline and the cohort's trendline are not significant, as the latter runs inside the confidence intervals of Braden's trendline.

In contrast, Cassie's FV-unit length shows a not very steep but incremental increase, although the positive correlation with time ($r = 0.33$) is not

FIGURE 2
FV-Units per Sentence Data and Trendlines: (a) Cohort Averages; (b) Braden with Cohort Trendline; (c) Cassie with Cohort Trendline



Note. FV/S = finite verb units per sentence.

significant ($p > 0.05$). It appears that this correlation is moderated by several oscillations in the data (Figure 3c): Cassie's trendline goes up from about 6 to 7 words per FV-unit over time, but the W/FV value jumps up to 8.5 at T6, to 9 at T9, and to 10.4 at T19, as well as drops to 4 at T17. Although Cassie's trendline has a visually steeper slope than the cohort trendline, this difference is not significant, as illustrated by the cohort trendline staying largely inside the confidence intervals of Cassie's trendline.

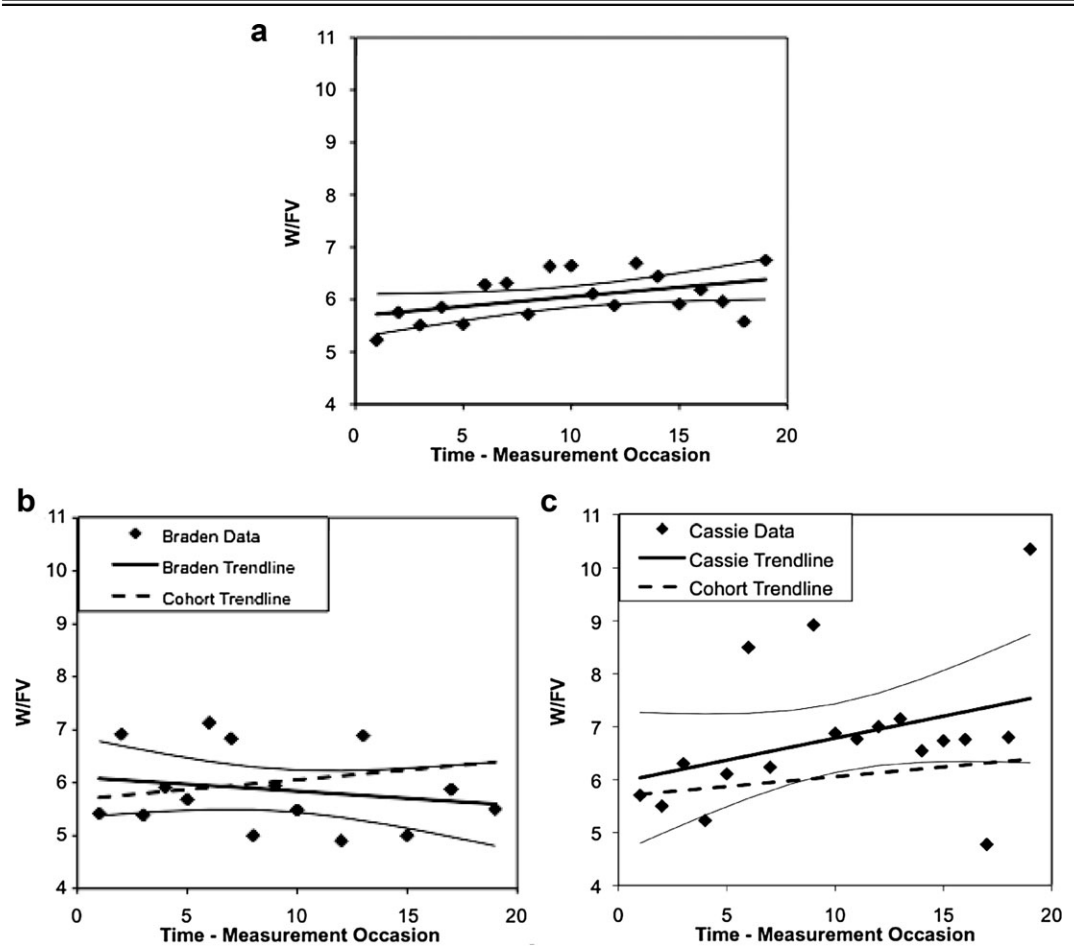
Complexification by Coordination

Figure 4a illustrates the cross-sectional cohort development of coordinating conjunctions (CC) as used per each 100 words. These values declined

with time, which is supported by a strong negative correlation ($r = -0.63$). The cohort trendline shows a diagonal decrease going down from about 5.8 to 3.5 CC per 100 words. One can also see a number of fluctuations reaching out beyond the confidence interval lines, which can in part be explained by differences in essay topics. For example, at T3, the task included describing family trees, which lends itself to listing multiple family members by using coordinated nominal phrases. This explains an upward fluctuation to 8 CC per 100 words.

The longitudinal data analysis shows that the cohort trendline runs inside the confidence intervals of both learners' trendlines (Figures 4b & 4c). However, the correlation of CC per 100 words with

FIGURE 3
Words per FV-Unit Data and Trendlines: (a) Cohort Averages; (b) Braden with Cohort Trendline; (c) Cassie with Cohort Trendline



Note. W/FV = words per finite verb units.

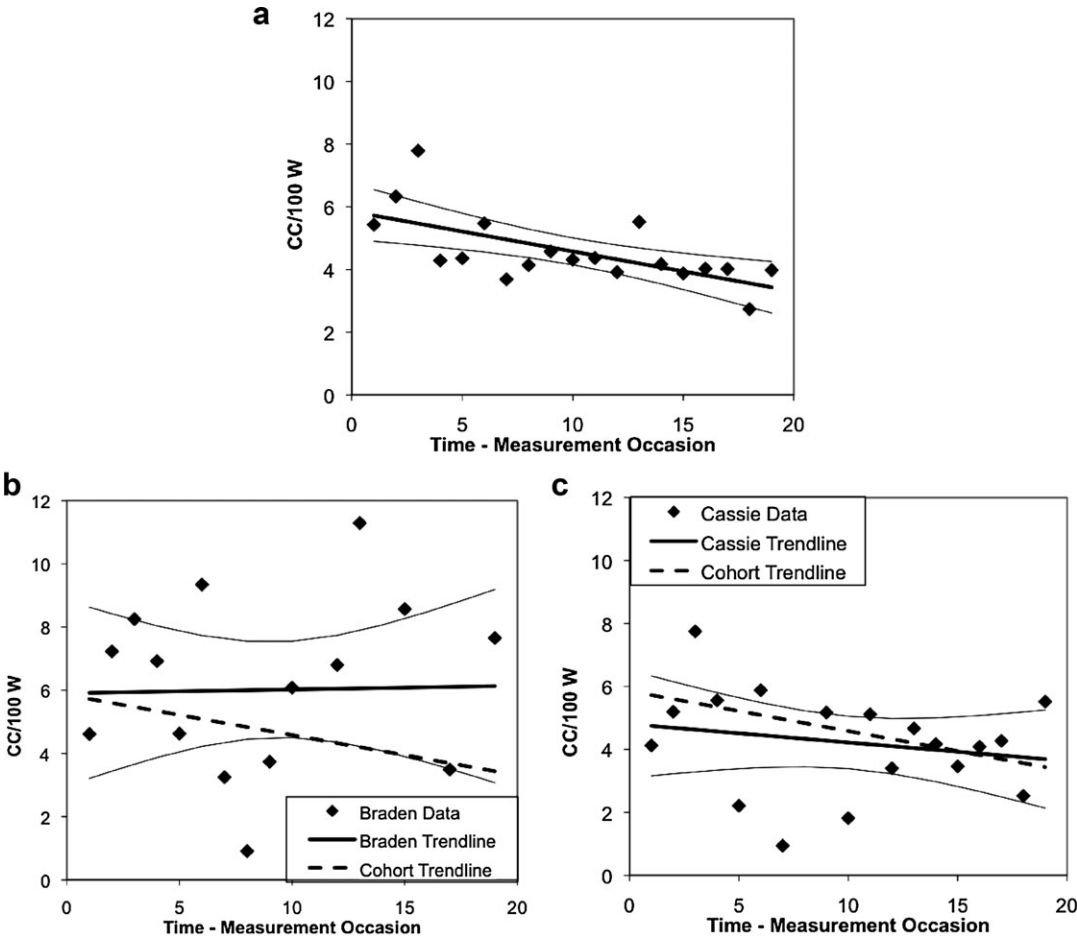
time was very weak and not significant for either Braden ($r = 0.07$) or Cassie ($r = -0.18$). This can be explained by the fact that the longitudinal data exhibit large variation: They are widely scattered around the trendlines. Braden's trendline is perfectly horizontal: His CC use does not significantly change over time and stays at 6 per 100 words on average. Cassie's trendline has a very slight decline from 4.8 to 4 words.

Complexification by Subordination

The cross-sectional cohort development of subordinating conjunctions (SC) as used per each 100 words is shown in Figure 5a. Focused instruction on subordination in general and SC in particular occurred at T8, and learners were not

expected to use these syntactic connectors prior to that point. However, the chart shows sporadic SC instances, even at earlier data points. This can be explained by the fact that participants in this study are adult, cognitively developed learners who eventually feel the need to express more complex thoughts and, therefore, to use more advanced textual cohesive devices. To do that, they may consult the dictionary or the teacher. The use of SC sharply jumps up from around 0 to 2.85 per 100 words at T8 due to the instructional focus, when learners were explicitly taught this grammatical feature and encouraged to use SC in their writing. The rate of SC use drops again at the next time point (T9) to 0.73 per 100 words but after that gradually increases with occasional fluctuations beyond the confidence intervals of the trendline.

FIGURE 4
CC per 100 Words Data and Trendlines: (a) Cohort Averages; (b) Braden with Cohort Trendline; (c) Cassie with Cohort Trendline



Note. CC/100 W = coordinating conjunctions per 100 words.

The correlation of this measure with time is positive and very strong ($r = 0.84$).

The longitudinal data analysis showed that both focal learners tried out subordinating conjunctions before focused instruction: Braden at T4 (Figure 5b) and Cassie at T5 (Figure 5c). Their SC use positively correlates with time, and the correlation is relatively strong ($r = 0.66$ for Braden, $r = 0.62$ for Cassie). Both learners' trendlines run below the cohort trendline; that is, their development lags behind the class average on this measure. However, the cohort trendline runs inside the confidence boundaries of Braden's trendline but above those of Cassie's trendline. This suggests that Braden's SC development is not significantly different from the class average, whereas Cassie's is. Additionally, both learners'

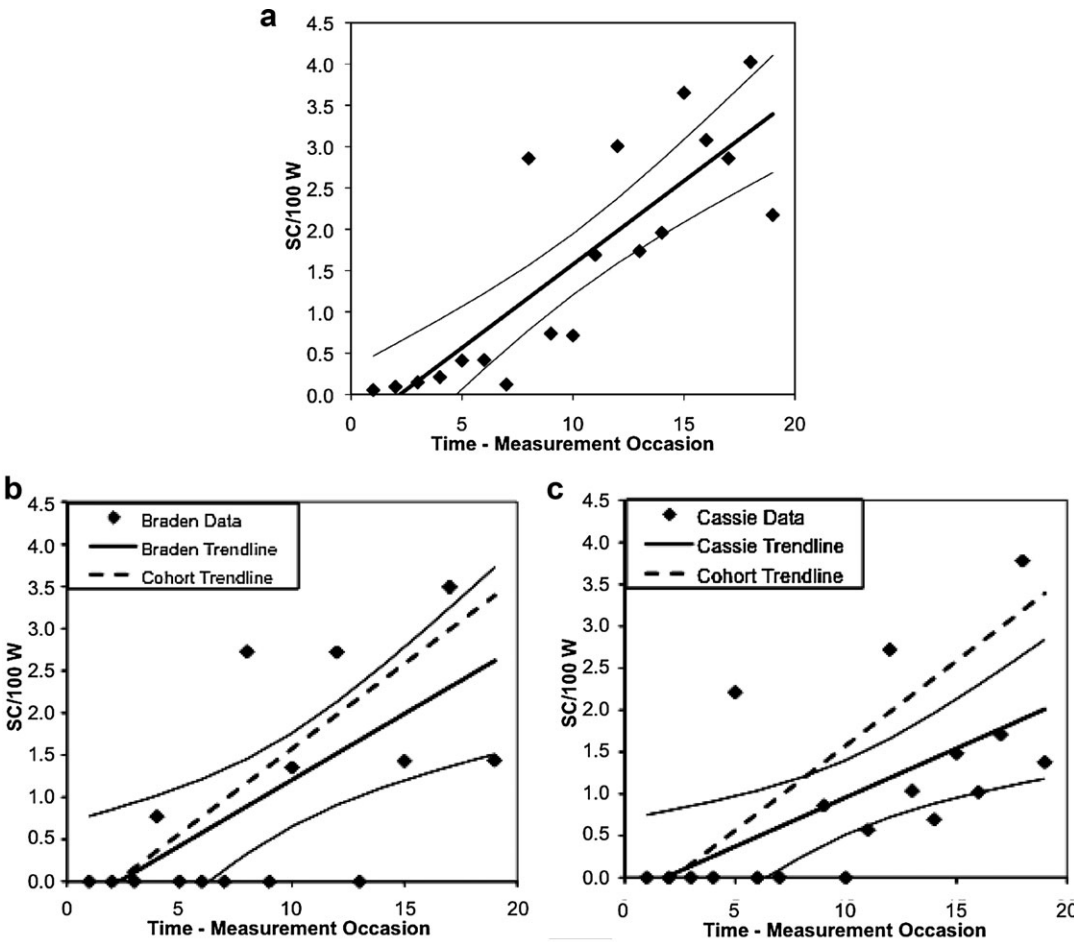
data are fairly scattered, indicating large variation, including occasional drops back to zero SC use.

Lexical Variety (CTTR)

As measured by the adjusted type-token ratio (CTTR), the cohort mean lexical variety showed a steady increase with only four data points falling outside the confidence intervals. The correlation of CTTR with time is very strong and positive ($r = 0.89$).

Both longitudinal learners develop their CTTR over time similarly to the class trendline: The correlation of this measure with time is strong and positive ($r = 0.59$ for Braden, $r = 0.60$ for Cassie). However, Figures 6a and 6b also show striking differences between them. Braden's CTTR starts

FIGURE 5
SC per 100 Words Data and Trendlines: (a) Cohort Averages; (b) Braden with Cohort Trendline; (c) Cassie with Cohort Trendline



Note. SC/100 W = subordinating conjunctions per 100 words.

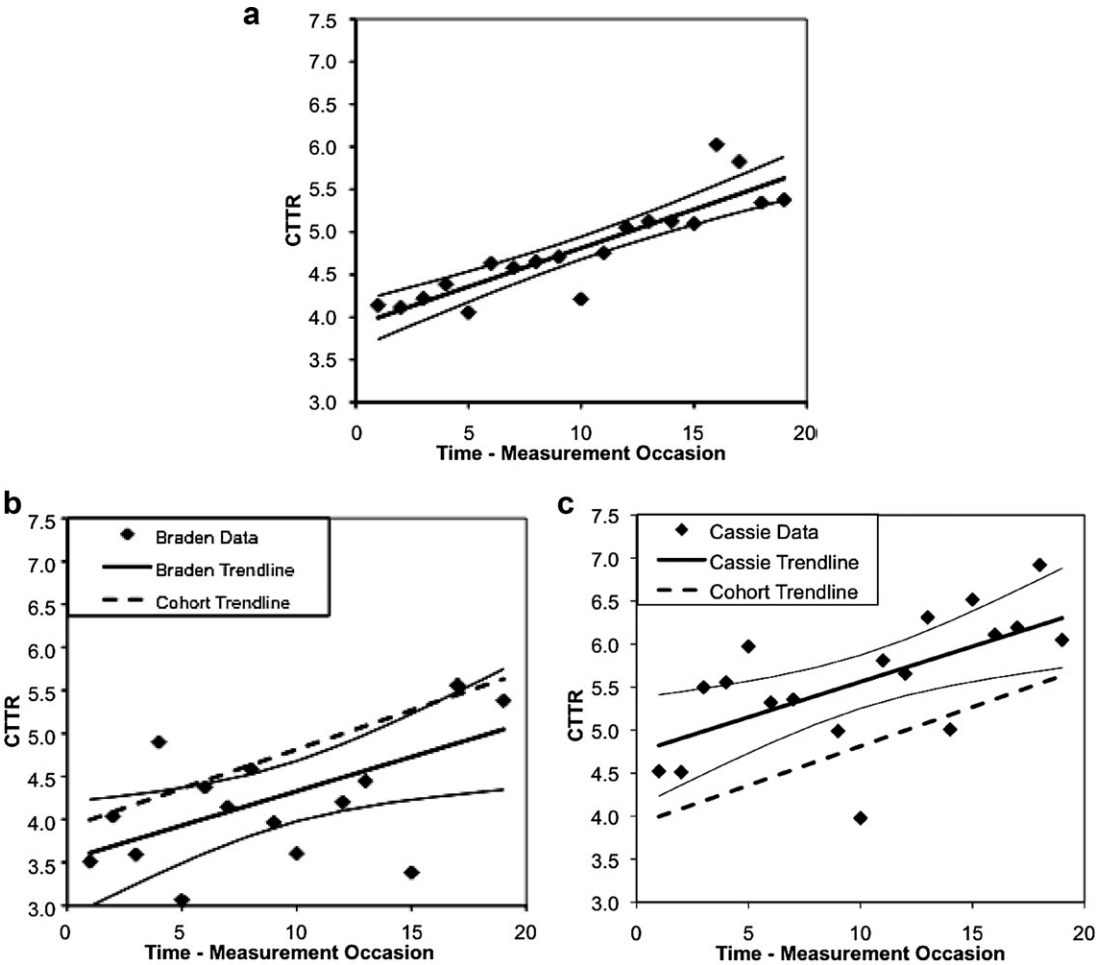
and continues to stay below the class average, with only two data points above the cohort trendline. Cassie's CTTR, in contrast, starts off above the class average and consistently surpasses it over the whole developmental course with only two data points below the cohort trendline. Moreover, the latter is consistently below the confidence intervals of Cassie's trendline, which shows that she is significantly above the class average on this measure.

RESULTS II: CORRELATION BETWEEN MEASURES

The analysis above showed that all measures for the cohort and most measures for the longitudi-

nal learners exhibited moderate to strong correlation with time. However, variation for specific measures was found between the two focal learners as well as between the longitudinal learners and the cohort. Additionally, since it was reasonable to assume that some specific measures were interrelated, further correlation analyses were conducted. Table 3 presents the correlation values between these selected complexity measures for the cohort averages as well as for the 2 focal learners. These results are discussed in the following sections. The final section also contains a qualitative analysis of the collocations of syntactic connectors in the longitudinal data.

FIGURE 6
CTTR Data and Trendlines: (a) Cohort Averages; (b) Braden with Cohort Trendline; (c) Cassie with Cohort Trendline



Note. CTTR = corrected type–token ratio.

General and Specific Complexity Measures

As shown in Figures 1–6, sentence length (SL), the most general complexity measure, correlated very strongly with time for both the cohort and longitudinal data. For the cohort data, this correlation translated into very similar (positive or negative) correlations of all more specific complexity measures (FV/S, CC, SC, CTTR) with both time and SL (cf. Tables 2 & 3). However, for the 2 focal learners, some intriguing differences were found. For Braden, SL correlated very strongly ($r = 0.81$) with FV/S and moderately yet insignificantly with CTTR ($r = 0.46$; $p > 0.05$).

For Cassie, SL correlated very strongly with CTTR ($r = 0.75$) and moderately yet significantly with both FV/S ($r = 0.54$) and W/FV ($r = 0.47$). The reader is also reminded that Braden surpassed the class average on the FV/S measure (Figure 2b) and Cassie on the CTTR measure (Figure 6c). This finding suggests that Braden’s general complexity (SL) mostly developed by increasing the number of FV-units per sentence, whereas Cassie predominantly made her sentences longer by using more words and more varied words. A closer analysis of Braden and Cassie’s use of coordinating and subordinating conjunctions (see next section) lends further support to this hypothesis.

TABLE 3
Correlations Between Selected Complexity Measures

		FV/S	W/FV	CC	SC	CTTR
Cohort	SL	0.88***	0.45	-0.59**	0.85***	0.79***
Braden		0.81***	0.16	0.57*	0.30	0.46
Cassie		0.54**	0.47*	-0.07	0.45	0.75***
Cohort	FV/S		-0.02	-0.58**	0.95***	
Braden			-0.44	0.33	0.63**	
Cassie			-0.45	-0.09	0.47	
Cohort	CC				-0.61**	
Braden					-0.36	
Cassie					-0.30	

Note. CC = coordinating conjunctions; CTTR = corrected type-token ratio; FV/S = finite verbs per sentence; SC = subordinating conjunctions; SL = sentence length. W/FV = words per finite verb unit.

* $p < 0.05$, ** $p < .01$, *** $p < 0.001$.

FV-Units per Sentence and Words per FV-Unit

The 2 focal learners developed differently vis-à-vis the two complexity measures based on FV-units: FV/S (Figures 2b & 2c) and W/FV (Figures 3b & 3c). Braden develops more on the former and Cassie on the latter. Furthermore, Braden significantly surpasses the cohort average development on the first measure, whereas Cassie's data do not significantly diverge from the cohort trendline on either.

Additionally, the scatterplots for Braden and Cassie suggested a curious pattern: It appeared that at some points, when the FV/S values showed sudden upward fluctuations, the W/FV values showed downward fluctuations, such as at T7, T9, and T15 for Braden (Figures 2b & 3b) and at T6, T9, T17, and T19 for Cassie. This finding prompted the researcher to test the correlation between the two complexity measures. The results showed a moderate negative correlation between FV/S and W/FV which, however, does not reach significance for either Braden ($r = -0.44$; $p = 0.08$) or Cassie ($r = -0.45$; $p = 0.06$). The cohort showed no correlation between these measures ($r = -0.02$). This difference between cross-sectional and longitudinal data illustrates how averaging may mask intra-individual correlation between features.

Coordination and Subordination

The data for syntactic connectors suggested an upward trend for SC and a downward trend for CC. Therefore, it was decided to measure the correlation between these two variables. The analysis confirmed a strong negative correlation between the strategies CC and SC for the cohort data ($r = -0.61$, $p < .01$). Additionally, an even

stronger correlation was found after T8 when the focused instruction on SC was administered ($r = -0.68$). For the two focal learners, a moderate yet insignificant correlation between CC and SC was found ($r = -0.36$ for Braden, $r = -0.30$ for Cassie). Interestingly, when measured after T8, this correlation became stronger for Braden ($r = -0.52$) but virtually disappeared for Cassie ($r = -0.17$).

Additionally, an extremely strong positive correlation was found between FV/S and SC for the cohort ($r = 0.95$). In contrast, FV/S correlates negatively with CC with a strong effect for the cohort ($r = -0.58$). This may indicate that on average, when learners started using more FV-units per sentence (which happened over time, see Figure 2a), these were typically subordinate and not coordinate clauses. This result was confirmed for SC ($r = 0.63$, $p < .01$) for Braden and approached significance for Cassie ($r = 0.47$; $p = 0.0515$). However, there was a quite different picture for CC: There was no significant correlation for either of the 2 focal participants. This finding highlighted another area of individual variability in comparison with class averages.

To shed more light onto the latter finding, a collocation analysis was performed for Braden and Cassie's use of CC and SC. Using the WordSmith Tools *Concord* function (Scott, 2008), concordance lines for all instances of CC and SC were retrieved and analyzed qualitatively. The results showed that Braden used a total of 28 SC tokens and Cassie a total of 37 SC tokens. However, it turned out that type-wise, both learners' data were similar: Each of them used 9 different types of SC conjunctions. The most frequent SC type by far was the complementizer *dass* ('that'): Each of the 2 participants used it 10 times. This is in line with

Sato (1988), who found that complement clauses were the most frequent type of subordination used by beginning learners. However, the 2 focal learners differed in the range of lexical items and structures governing the complementizer *dass*. Braden used *dass* in a fixed construction *ich dachte, dass* ('I thought that') in 5 instances out of 10. In contrast, all of Cassie's 10 *dass* instances occur in collocations with different verb lexemes or forms (e.g., *er hatte erklärt, dass* 'he had explained that'; *deshalb hat er gedacht, dass* 'therefore, he thought that'; *ich glaube, dass* 'I think that'; *Ich hoffe, dass* 'I hope that'; *Ich würde nicht sagen, dass* 'I would not say that'), thus exhibiting much richer lexicogrammatical variety and supporting the findings of the earlier CTTR analysis.

The CC concordance analysis showed that both learners predominantly used the conjunctions *und* ('and') and *aber* ('or'). Braden used 95 *und* and 16 *aber*; Cassie used 89 *und* and 20 *aber*. A closer look at the collocations of the most frequent conjunction *und*⁷ revealed that both learners used it to connect FV-units only in about a third of all occurrences (Example 1). In contrast, two thirds of *und* instances connected different kinds of paratactic constructions (Examples 2 & 3), thus contributing to lengthening of FV-units rather than to increasing the number of coordinated FV-units.

EXAMPLES

- (1) *Am Morgen, ich frühstücke und Kaffee trinken.*
(FV-units)
'In the morning, I have breakfast and drink coffee.'
- (2) *Meine Tante Kay und ihr Mann Bill haben zwei Kinder und zwei Enkel.* (noun phrases)
'My aunt Kay and her husband Bill have two children and two grandchildren.'
(noun phrases)
- (3) *Die Stadt ist größer und schneller.* (predicative adjectives)
'The city is bigger and faster.'

SUMMARY AND DISCUSSION

Analysis of the development of syntactic complexity shows a general upward trend in both the cohort and longitudinal data on most of the measures used in this study: a general complexity measure (sentence length) and more specific complexity measures (finite verb units, subordination, and lexical variety). In other words, learners write more complex essays as they progress from the novice level through a college-level instruc-

tional sequence of four semesters. This overall result corroborates the assurance by Larsen-Freeman (2006) that "[f]oreign language teachers and learners can take heart" because over time, learners' "writing has become more complex in grammar and in vocabulary" (p. 598). This finding was confirmed for both timed and untimed essay writing. Although students' development in writing complexity is more apparent when they are allowed to write without time pressure, the developmental trend still holds even when only timed essays are considered. More specific results are discussed next in comparison with previous research.

Cohort Data

General and Specific Syntactic Measures. The findings for sentence length triangulate the results from Byrnes (2009), Byrnes et al. (2010), and Cooper (1976), whose participants, like in this study, were American college-level learners of German. The cited studies showed an incremental increase in means of generic complexity measures (words per sentence and/or words per T-unit) at transition points between curricular levels. This study supported and augmented these findings by the correlation analysis of densely plotted data, which showed a steady linear increase of the SL trendline with tight confidence intervals (Figure 1a). This finding presents evidence that learners in this study consistently increased their sentence length during each semester, and not only between semesters. A similar result was achieved for sentence length in clause-type units (FV-units/S), which parallels the trend of a significant increase in the number of clauses per T-unit shown between adjacent levels I to III in Byrnes et al. (2010).⁸ The data collection in this study ended after the fourth semester (about 255 contact hours total), which roughly corresponds to the end of level III in Byrnes et al. (2010). Therefore, no data are available to test whether the number of clause-type units per a generic syntactic unit declines at higher proficiency levels (as found in the cited studies).

In contrast, the picture for the FV-unit length in words is different. Byrnes (2009) and Byrnes et al. (2010) found that clause length was not significantly different between levels I and II but increased with each next level. Based on this finding, an increase in FV-unit length could have been expected in the fourth quartile of this study's timeline, but this was not substantiated in the data (Figure 3a). In this regard, the participants in this study seem more similar to those of Cooper (1976), who did not see an increase in clause

length until the fourth year of study (Cooper provides no data for the first year). An explanation of this fact may lie in the differences in the instructional policies and practices. Whereas the curriculum in this study's instructional setting generally switched from more narrative to more expository writing at the focal juncture (fourth semester), the GUGD curriculum, apparently, had a much more explicit focus on secondary public discourse genres and associated linguistic features such as nominalization and other means of clause extension at the corresponding level III. Indeed, as Norris and Ortega (2009) explain, "mean length of clause is radically different from the other length-based measures" because

any increases can only result from the addition of pre- or postmodification within a phrase (via adjectives, adverbs, prepositional phrases, or nonfinite clauses) or as a result of the use of nominalizations, or the process of reduction of clauses into phrases which help to condense information. (p. 561)

Therefore, increasing clause length may serve as an indicator of advanced levels of L2 proficiency ascertained at level III for GUGD students but not in this study or Cooper's (1976) study.

Syntactic and Lexical Measures. The study showed that on average, this cohort's lexical variety (CTTR) positively and very strongly correlates with sentence length and with time. Interestingly, the strong correlation between SL and CTTR for both the cohort and the longitudinal data in this study is the reverse of the pattern found by Verspoor et al. (2008) in the written data of an advanced learner of English, whose SL and TTR development correlated negatively. However, the data for that same learner showed a weak to moderate positive correlation at the beginning of the data collection period. The present study thus provides support for Verspoor et al.'s conclusion that varied word use and the length of sentences have a complex relation to each other that "changes dynamically over time" (p. 225). At the beginning stages of L2 proficiency, as is the case for the learners in this study, these two variables may act as "connected growers" (van Geert, 1994) that support each other's development. Indeed, beginning learners who need to increase the length of their initially very basic sentences, need not only new syntactic structures but also new vocabulary to fill their sentences. In Verspoor et al.'s words, sentence length "reflects to a great extent vocabulary acquisition and the ease with which vocabulary is used" (2008, p. 220). In contrast, at more advanced stages of L2 development, lexical and syntactic

complexity may enter into a competitive relationship, as shown by Verspoor et al.

Coordination and Subordination. In contrast, CC and SC have been shown to correlate negatively in this data. Moreover, CC has also been shown to negatively correlate with time, sentence length, and the number of FV-units per sentence (Tables 2 & 3). This finding can be explained as follows. At the very outset of language study, an overwhelming majority of syntactic connectors used by learners are coordinating conjunctions (Figure 4a). This is not surprising because coordination is the only means of syntactic complexification that students are taught during the first semester of study. After focused instruction on subordinating conjunctions (T8), the use of CC gradually decreased and the use of SC gradually increased (Figures 4a & 5a), which also goes hand in hand with the increase of (subordinate) FV-units (Table 3). This finding corroborated the results from previous studies of beginning language learners (Bardovi-Harlig, 1992; Ishikawa, 1995). However, at some observation points, the pattern was reversed: CC went up at T9, T13, and T19, whereas SC went down. This reverse effect may be explained by the task effect: For example, essay 19 was written in class, and learners predictably used fewer cognitively challenging SC (than in essay 18, which was written at home) and resorted back to using more CC. More dynamic research methods are needed to account for these interactions between factors.

Moreover, the linear increase of SC in this study can be loosely paralleled to the increase of clauses per T-unit (a different subordination measure) found in Byrnes et al. (2010) through levels I to III. On the one hand, this similarity may be explained by the gradually increasing focus on reasoning components in both writing programs, which require the use of subordinating conjunctions (Byrnes et al., 2010; Michel, 2010). On the other hand, the results regarding CC should be interpreted with caution. Coordination is a much more heterogeneous linguistic feature than SC because it can combine components at different syntactic levels (words, clauses, phrases), as shown by the analysis of concordance lines with the CC *und*. Moreover, coordination may indicate lesser as well as greater syntactic complexity (e.g., in the case of conjoined clauses, see Robinson, 2007). In fact, Cooper (1976) found mixed results regarding coordinated structures: They declined from level II to III in his learners but gradually increased again with each next proficiency level. A detailed analysis of coordination in the writing of the 2 focal

learners was performed in a separate study (see Vyatkina, in press).

Longitudinal Data

The longitudinal data analysis richly illustrates the variability in developmental pathways of 2 individual learners vis-à-vis the backdrop of the cross-sectional developmental trend. The analysis of sentence length shows that Braden and Cassie's developmental pathways oscillate closely around the average class trendline on this generic syntactic complexity measure. It was therefore interesting to track how these two learners developed based on more specific complexity measures; that is, what are the specific means by which their sentence length increases? The results show that Braden makes his sentences progressively longer primarily by adding more FV-units. He is consistently and significantly above the class average on this measure. In contrast, Cassie's sentence length most strongly correlates with CTTR, and she consistently and significantly surpasses her classmates on this measure. The qualitative analysis confirms this hypothesis, at least, for SC collocations: Cassie uses the SC *dass* with a variety of words and constructions. This variation is in stark contrast with Braden's use of the same connector mostly in a fixed construction *ich dachte, dass*. The specific verb form *dachte* as used by Braden in conjunction with *dass* is an example of chunk-type word combinations, or "lexical entry points into complex structures" typical of beginning learners, according to Sato (1988, p. 392). This observation points to the fact that Cassie may have experimented with a richer repertoire of lexicogrammatical features typical of more advanced learners. This assumption finds further support in an observation of Cassie's instructor who praised this student's creativity in her essays.

Additionally, Cassie's sentence length increase moderately correlated with an increase in the FV-unit length. As discussed earlier, lengthening of clause-type units is generally considered a more advanced complexification strategy as opposed to the increase of the number of clause-type units per sentence (as used by Braden). Therefore, a positive correlation between Cassie's SL and FV-unit length may suggest that she has reached a higher stage of syntactic complexification than her classmates. Indeed, a separate qualitative study of the 2 focal learners confirmed this assumption, showing that Cassie used a wider variety of more advanced syntactic constructions (such as infinitives and participles) than Braden (Vyatkina, in press).

Finally, a moderate negative correlation between FV-units per sentence and words per FV-unit was found for the 2 focal learners, whereas the cohort data showed no correlation. While this finding suggests a hypothesis about a competitive relationship between the frequency and length of clause-type units in these individuals' longitudinal data, nonlinear dynamic methods will be needed to support or refute it.

IMPLICATIONS AND CONCLUSION

This study analyzed the development of writing complexity in college-level learners of German over four semesters of study beginning at the novice level. Various lexicogrammatical measures of complexity were used, and their correlations were explored cross-sectionally and longitudinally. Comparisons were made between cross-sectional trendlines and individual development paths taken by 2 learners. The results confirm that learners follow some general developmental trends established in previous CAF research, but considerable variability was also illustrated and explained. In this way, this research responds to Byrnes's (2009) call to study both individual values and group values and "to find regularities, but to find them in variations that are themselves specified" (p. 63). The relationships between different complexity measures were also explored.

Furthermore, this study contributes to the methodology for conducting developmental SLA research. First, selected CAF measures were described in detail, which will enable future applications of the study design to other contexts. The study shows that frequencies of automatically annotated POS such as finite verbs and syntactic connectors can be used as convenient proxy measures for syntactic complexity. Second, the study showed how cross-sectional and longitudinal data can be used as complementary data sources yielding interilluminating research results. In particular, using scatterplots along with trendlines and confidence interval lines illustrates cross-sectional to longitudinal comparisons.

The results were compared to a series of findings from previous research conducted for similar learner populations (Byrnes, 2009; Byrnes et al., 2010; Cooper, 1976). The cited studies as well as this study have shown that the general syntactic complexity and the amount of subordination linearly increase in the writing of beginning to intermediate learners. This study has additionally shown that the linear increase trend held both for time points T1 to T10, at which modifications of a narrative task were used, and

for time points T11 to T19, at which a reasoning component was added. This is an important methodological finding confirming that these measures (sentence or T-unit length in words and clause-type units as well as normalized counts of subordinating conjunctions) may be used as reliable indices of progressing from beginning to intermediate L2 proficiency, which are sufficiently resistant to differences in instructional settings and tasks (the latter being a particularly severe threat to validity in longitudinal designs).

In contrast, the clause-type unit length in words did not work as a developmental unit in this study, at least for the cohort data, because the values did not change significantly over time. This finding pointed to curricular differences between this study's program and the program described by Byrnes (2009) and Byrnes et al. (2010), who showed a significant increase of clause length (a characteristic of advanced L2 capacities) at a comparable time juncture in terms of the number of instructional hours. Finally, this study has shown that the general amount of coordination correlates negatively with the amount of subordination, although it was pointed out that coordination is a potentially heterogeneous measure and merits a further exploration.

It has to be noted that the complexity measures and methods in this study have inevitable limitations, which show directions for future research. First, the decision to use only automatically computed measures brings up the issue of software accuracy. This was evaluated here by spot-checks; however, future investigations should (ideally) follow a semi-automatic approach with a systematic correction of tagger errors (Garretson & O'Connor, 2007, see also Vyatkina, in press). Second, to enhance research comparability, more studies are needed with data manually annotated for syntactic complexity measures used for similar learner populations (such as clauses, T-units, and lexical sophistication). Third, this study has only begun to tap into the development of more specific complexity measures that present a rich potential for future research. More studies are already underway that explore a large set of POS-based complexity measures in response to specific writing tasks for a focal cohort and individuals. Fourth, this study used linear correlation methods, which was considered appropriate as a first approximation for comparing cross-sectional and individual data. However, a number of longitudinal findings yielded hypotheses that may be tested with nonlinear dynamic methods (Verspoor et al., 2011), such as (potentially) supporting relationship between sentence length

and lexical variety or a competing relationship between coordination and subordination (at a certain point in the instructional progression).

On a final note, this study has a potential to be expanded with analyses of data collected for subsequent learner cohorts. Such analyses may be particularly revealing because of the changes in the curricular approach to writing implemented in the focal program over the years. In particular, efforts have been made to enhance the ecological validity (Byrnes et al., 2010) of the writing component (see Vyatkina, 2011). Future comparisons of the learning outcomes for different learner cohorts may validate such curricular changes, which would have implications for both writing pedagogy and research in L2 education.

NOTES

¹ Byrnes et al. (2010) explain that they designed this BWT for students in their program as "a task that (a) could be reasonably attempted by learners at four broadly different ability levels; (b) could be completed within a relatively short amount of time; and (c) could, nevertheless, elicit a reasonably trustworthy indication of learners' writing abilities and syntactic patterns" (p. 165). Given that students in this study engaged in discussions of German short stories for a whole semester, writing a short book review was considered an appropriate semester-final writing task.

² Although some researchers account for productivity, or text length in words, in complexity analyses, it is widely accepted that it is rather a separate CAF measure akin to fluency (Polio, 2001). Following this assumption, productivity is not analyzed in this study. However, as additional information on the study background, Table 1 demonstrates that the average essay length progressively increases (with some upward and downward oscillations) from 68 words at T1 to 161 words at T19 (which was expected). Also, learners wrote longer essays under untimed conditions (which was also expected). Furthermore, it should be noted that whereas Braden's essay length mostly fluctuated around the cohort trendline, Cassie frequently wrote longer and several times much longer essays than the class average.

³ Although clauses are rarely defined explicitly, the following differences emerge from discussion sections and/or examples in some recent studies: Norris and Ortega (2009) discuss finite clauses but consider nonfinite verb constructions phrases, i.e., subclausal elements (p. 561, see also Lu, 2011); Byrnes et al. (2010) count infinitival constructions *zu* ('to') + *infinitive* as nonfinite clauses but infinitives governed by modal verbs as subclausal elements (p. 168); whereas Kuiken, Vedder, and Gilabert (2011) consider all nonfinite verb constructions governed by modal verbs separate clauses (at least for some Germanic and Romance languages, following Housen, 2002, p. 106).

⁴ Another study that reports on specific cases of coordination and subordination, including modification, for the 2 longitudinal learners accounts for this limitation (Vyatkina, in press).

⁵ Using an automatic POS tagger trained on native data for learner data implies that only native-like forms were calculated, i.e., accuracy was implicitly taken into account although it is not the focus of this study. It should be noted that using an automatic tool always assumes a certain percentage of annotation errors (Granger, 2002; Meurers & Müller, 2009). The tagger accuracy was evaluated by two independent raters and estimated at about 96%, which is fairly high. Conjunctions are short uninflected words in which learners make few spelling mistakes, and many misspelled verb forms such as *arbiere* instead of *arbeite* were recognized as FV by the tagger. It must be added that human annotation is not free of errors either, and automatic taggers' errors are at least systematic.

⁶ Sentence length is expected to further increase with growing L2 proficiency based on Cooper's (1976) data, which showed consistent growth with each year of study and a significant increase after every 2 years.

⁷ This observation was also confirmed by the cross-sectional wordlist analysis: The word *und* is by far the most frequent word, not only the most frequent connector, in this learner corpus. Moreover, this pattern parallels native speaker usage of *und* that has been shown to be one of the most frequent words in German (Tschirner, 2005).

⁸ No direct comparisons are possible between the units used in this study (automatically annotated FV-units) and the cited studies (manually annotated finite and nonfinite clauses in Byrnes et al., 2010, and finite clauses in Cooper, 1976). However, general trends can be compared because all these units are similar in the sense that they measure syntactic complexity at the level lower than a sentence but higher than a word and have a verb at their core.

ACKNOWLEDGMENTS

This study was supported in part by the University of Kansas General Research Fund allocations Nos. 2302139 and 2301446. I would like to acknowledge Emily Hackmann for her help in evaluating the tagger accuracy and Jonathan J. Van Tassel for assistance with correlation analysis. I would also like to thank William J. Comer and the anonymous reviewers for their comments on earlier drafts of this article.

REFERENCES

- Aarts, J., & Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 132–141). London: Longman.
- Abrams, Z. (2010). Writing. In C. Blyth (Ed.), *Foreign language teaching methods*. Texas Language Technology Center, University of Texas at Austin. Retrieved from <http://coerll.utexas.edu/methods>
- Arthur, B. (1979). Short-term changes in EFL composition skills. In C. Yorio, K. Perkins, & J. Schachter (Eds.), *On TESOL '79: The learner in focus* (pp. 330–342). Washington, DC: TESOL.
- Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly*, 26, 390–395.
- Briggs, J., Di Donato, R., Clyde, M., & Vansant, J. (2008). *Workbook to accompany Deutsch, Na Klar!: An introductory German course*. New York: McGraw-Hill.
- Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, V. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). Philadelphia/Amsterdam: John Benjamins.
- Byrnes, H. (2009). Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. *Linguistics and Education*, 20, 50–66.
- Byrnes, H., Maxim, H., & Norris, J. M. (2010). Realizing advanced foreign language writing development in collegiate education: Curricular design, pedagogy, assessment [Monograph]. *Modern Language Journal*, 94(s1).
- Byrnes, H., & Sinicrope, C. (2008). Advancedness and the development of relativization in L2 German: A curriculum-based longitudinal study. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 109–138). New York: Routledge/Taylor & Francis.
- Casanave, C. (1994). Language development in students' journals. *Journal of Second Language Writing*, 3, 179–201.
- Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *Journal of Educational Research*, 69, 176–183.
- de Bot, K., Lowie, W., & Verspoor, M. H. (2011). Introduction. In M. H. Verspoor, K. de Bot, & W. Lowie (Eds.), *A dynamic approach to second language development. Methods and techniques* (pp. 1–4). Philadelphia/Amsterdam: John Benjamins.
- Di Donato, R., Clyde, M., & Vansant, J. (2008). *Deutsch, na klar!: An introductory German course*. Boston: McGraw-Hill.
- Draper, N. R., & Smith, H. (2001). *Applied regression analysis*. New York: Wiley.
- Ellis, N., & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics—introduction to the special issue. *Applied Linguistics*, 27, 558–589.
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.
- Garretson, G., & O'Connor, M. C. (2007). Between the Humanist and the Modernist: Semi-automated analysis of linguistic corpora. In E. Fitzpatrick (Ed.), *Corpus linguistics beyond the word: Corpus*

- research from phrase to discourse (pp. 87–106). Amsterdam:
- Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Philadelphia/Amsterdam: John Benjamins.
- Granger, S., Kraif, O., Ponton, C., Antoniadis, G., & Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL*, 19, 252–268.
- Halliday, M. A. K., & Matthiessen, C. M. I. M. (1999). *Construing experience through meaning: A language-based approach to cognition*. London: Continuum.
- Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1, DOI: 10.1017/S2041536210000103.
- Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly*, 18, 87–107.
- Housen, A. (2002). Second language development in the European School system of multilingual education. In D. W. C. So & G. M. Jones (Eds.), *Education and society in plurilingual contexts* (pp. 96–128). Belgium, Brussels: VUB University Press.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30, 461–473.
- Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (NCTE research report No. 3). Champaign, IL: National Council of Teachers of English.
- Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, 4, 51–69.
- Kern, R. G., & Schultz, J. M. (1992). The effects of composition instruction on intermediate level French students' writing performance: Some preliminary findings. *Modern Language Journal*, 76, 1–13.
- Kuiken, F., Vedder, I., & Gilabert, R., (2011, March). *Syntactic complexity in L2 writing as an indicator of L2 proficiency*. Paper presented at the Annual Conference of the American Association for Applied Linguistics, Chicago, IL.
- Larsen-Freeman, D. (1983). Assessing global second language proficiency. In H. Seliger & M. Long (Eds.), *Classroom-oriented research in second language acquisition* (pp. 287–304). Rowley, MA: Newbury House.
- Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27, 590–619.
- Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in Second Language Acquisition. *Applied Linguistics*, 30, 579–589.
- Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45, 36–62.
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal*, 96, 190–208.
- Maxim, H., (2011, March). *Establishing a curricular trajectory: A socio-semiotic perspective on text selection and sequencing*. Paper presented at the Annual Conference of the American Association for Applied Linguistics, Chicago, IL. Retrieved from <http://userwww.service.emory.edu/~hmaxim/presentations.html>
- Meurers, W. D., & Müller, S. (2009). Corpora and syntax. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 920–933). Berlin: Mouton de Gruyter.
- Michel, M. (2010). *Cognitive and interactive aspects of task-based performance in Dutch as a second language (Unpublished doctoral dissertation)*. University of Amsterdam, The Netherlands.
- Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30, 555–578.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24, 492–518.
- Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szendrői (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 127–155). Berlin: Mouton de Gruyter.
- Ortega, L., & Byrnes, H. (2008). The longitudinal study of advanced L2 capacities: An introduction. In L. Ortega & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 3–20). New York: Routledge/Taylor & Francis.
- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30, 590–601.
- Polio, C. (2001). Research methodology in second language writing research: The case of text-based studies. In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 91–115). Mahwah, NJ: Lawrence Erlbaum.
- Ravid, D. (2005). Emergence of linguistic complexity in later language development: Evidence from expository text construction. In D. D. Ravid & H. B. Shyldkrot (Eds.), *Perspectives on language and language development: Essays in honor of Ruth A. Berman* (pp. 337–356). London: Kluwer Academic.
- Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics*, 45, 237–257.
- Ryshina-Pankova, M. (2010). Toward mastering the discourses of reasoning: Use of grammatical-

- metaphor at advanced levels of foreign language acquisition. *Modern Language Journal*, 94, 181–197.
- Sato, C. (1988). Origins of complex syntax in interlanguage development. *Studies in Second Language Acquisition*, 10, 371–395.
- Saville, N. (2010). The English profile programme: Background, current issues and future prospects. *Language Teaching*, 43, 238–244.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK. Retrieved from <http://www.ims.uni-stuttgart.de/~schmid/>
- Scott, M. (2008). WordSmith Tools (Version 5) [Computer software]. Liverpool, UK: Lexical Analysis Software.
- Skehan, P. (1989). *Individual differences in second language learning*. London: Edward Arnold.
- Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics*, 31, 532–553.
- Teichert, H., & Teichert, L. (2005). *Allerlei zum Lesen* [All kinds of reading]. Boston: Houghton Mifflin Co.
- Tschirner, E. (2005). Korpora, Häufigkeitslisten, Wortschatzerwerb [Corpora, frequency lists, vocabulary acquisition]. In A. Heine, M. Hennig, & E. Tschirner (Eds.), *Deutsch als Fremdsprache—Konturen und Perspektiven eines Fachs* [German as a foreign language—contours and perspectives of a discipline] (pp. 133–149). München, Germany: Iudicium.
- van Geert, P. (1994). *Dynamic systems of development: Change between complexity and chaos*. New York: Harvester.
- Verspoor, M. H., de Bot, K., & Lowie, W. (Eds.). (2011). *A dynamic approach to second language development. Methods and techniques*. Philadelphia/Amsterdam: John Benjamins.
- Verspoor, M., Lowie, W., & van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *Modern Language Journal*, 92, 214–231.
- Verspoor, M., & Sauter, K. (2000). *English sentence analysis: An introductory course*. Philadelphia/Amsterdam: John Benjamins.
- Vyatkina, N., (in press). Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus. *Modern Language Journal*, 97(s1).
- Vyatkina, N. (2011). Writing instruction and policies for written corrective feedback in the basic language sequence. *L2 Journal*, 3, 63–92.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Second Language Teaching & Curriculum Center, University of Hawai'i at Manoa.

APPENDIX

Instruction-Embedded Elicitation Tasks (based on Briggs, Di Donato, Clyde, & Vansant, 2008)

1. Who are you?
2. Your apartment, your friends
3. Your family
4. Your daily routine
5. Description: Your favorite clothes OR Your lucky charm OR A shopping day
6. Party plan
7. Your last weekend (sequence of activities in the present perfect tense)
8. You as a human (personal characteristics) OR a human (and human life)
9. Your town
10. Comparison of two trips (in the present perfect tense)
11. The life in the future (including a consequential explanation)
12. Describe yourself as a person or a person who you respect (including a consequential explanation)
13. You and the media OR You and technology (including a consequential explanation)
14. A (societal or personal) problem / challenge (including a possible solution)
- 15, 16, 17, 18. Interpretation. Use your imagination and think about the deeper meaning of this story (3–7 more specific prompts provided)
19. Book review. Choose a book, either fictional or nonfictional, that you have read and write an article for a regular feature in a student newspaper (a few more specific prompts provided)